

# A Realistic Simulation Framework for Analog/Digital Neuromorphic Architectures

**Fernando M. Quintana<sup>1,2,3†</sup>, Maryada<sup>4</sup>, Pedro L. Galindo<sup>3</sup>, Elisa Donati<sup>4</sup>, Giacomo Indiveri<sup>4</sup>, Fernando Perez-Peña<sup>3</sup>**

<sup>1</sup> Bio-Inspired Circuits and Systems Lab, Zernike Institute for Advanced Materials, University of Groningen, Netherlands

<sup>2</sup> Groningen Cognitive Systems and Materials Center, University of Groningen, Netherlands

<sup>3</sup> School of Engineering, University of Cádiz, Puerto Real, Cádiz, Spain

<sup>4</sup> Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland

E-mail: f.m.quintana.velazquez@rug.nl

October 2024

**Abstract.** Developing dedicated mixed-signal neuromorphic computing systems optimized for real-time sensory-processing in extreme edge-computing applications requires time-consuming design, fabrication, and deployment of full-custom neuromorphic processors. To ensure that initial prototyping efforts exploring the properties of different network architectures and parameter settings lead to realistic results, it is important to use simulation frameworks that match as best as possible the properties of the final hardware. This is particularly challenging for neuromorphic hardware platforms made using mixed-signal analog/digital circuits, due to the variability and noise sensitivity of their components. In this paper, we address this challenge by developing a software spiking neural network simulator explicitly designed to account for the properties of mixed-signal neuromorphic circuits, including device mismatch variability.

The simulator, called **ARCANA (A Realistic Simulation Framework for Analog/Digital Neuromorphic Architectures)**, is designed to reproduce the dynamics of mixed-signal synapse and neuron electronic circuits with autogradient differentiation for parameter optimization and GPU acceleration. We demonstrate the effectiveness of this approach by matching software simulation results with measurements made from an existing neuromorphic processor. We show how the results obtained provide a reliable estimate of the behavior of the spiking neural network trained in software, once deployed in hardware. This framework enables the development and innovation of new learning rules and processing architectures in neuromorphic embedded systems.

**Keywords:** SNN, DPI, neuromorphic, PyTorch, DYNAP-SE

## 1. Introduction

Mixed-signal neuromorphic circuits emulate the neural and synaptic dynamics observed in real neural systems, sharing similarities such as limited precision, heterogeneity, and high sensitivity

† Corresponding author

to noise [4, 14], which are often overlooked in Artificial Intelligence (AI) workloads. Typically, software simulations of Spiking Neural Network (SNN) use bit-precise activation functions identical for all neurons in a network with high-resolution parameters lacking biologically realistic details. Conversely, tools used in computational neuroscience, such as “Neuron” and “NEST”, capture bio-realistic properties of neurons and synapses—such as ion channels and neuronal morphology [7, 8, 25]. In other words, these simulators attempt to replicate real hardware—in this case, *the brain*—but do not offer modules for hyper-parameter optimization. Along similar lines, the neuromorphic community could benefit from a realistic simulator for electronic circuits that emulates the biophysics of neurons and synapses as well as the non-idealities of their analog substrates.

In this paper, we introduce a PyTorch-based [17] simulation platform, *ARCANA*, for mixed-signal neuromorphic systems. *ARCANA* is optimized to simulate neurons and synapses using differential equations derived from their electronic circuits [1]. This seamless integration provides an easy-to-use infrastructure and libraries for parameter optimization, thereby facilitating the development of neural networks deployable on neuromorphic hardware.

This enables neuromorphic engineers, computational neuroscientists and AI application developers to test and validate SNN architectures in a fast prototyping environment that abstracts the underlying hardware intricacies [11, 18, 21, 24, 27].

A hardware-aware network can be effortlessly deployed on an inference chip, such as those from the DYNAP family [16, 22]. Moreover, the simulator can be readily adapted to incorporate hardware-specific constraints during model training, enabling a unified framework across various platforms. This approach offers two significant advantages: (1) it simplifies the transfer of a single model across multiple platforms, ensuring seamless cross-platform deployment, and (2) it facilitates benchmarking of different hardware using diverse datasets. Additionally, if a chip supports on-chip learning, deploying a pre-trained network enables continuous adaptation to new real-world data.

## 2. Methods

The mixed-signal circuit equations used in this framework are based on the Differential Pair Integrator (DPI) circuit [1], which is commonly employed to implement both analog synapse and neuron circuits [4, 12, 15, 19, 23, 26]. The DPI synapse equations faithfully reproduce realistic synaptic dynamics [1] and the corresponding neuron equations mimic the behavior of the Adaptive Exponential Integrate & Fire (AdExpI&F) neuron model [2]. Because the DPI is a current-mode log-domain filter, both the synapse and the internal state variables of the neurons are expressed as *currents* (e.g. membrane current), even though theoretical and computational models, typically represent the neuron’s state as the membrane *potential*. In the simulator, we adopt terminology and variable names that closely match those used in the electronic circuits.

The mixed-signal neuromorphic processor used to validate the simulations is the Dynamic Neuromorphic Asynchronous Processor (DYNAP-SE) [16]. It is a multi-core SNN processor with four cores, each with 256 Adaptive exponential Leaky Integrate-and-Fire (AdexLIF)

neurons, providing a total of 1024 neurons [16].

### 2.1. DPI synapse

The simulator models four different types of synapses: two excitatory synapses (AMPA and NMDA) and two inhibitory (GABA<sub>A</sub> and GABA<sub>B</sub>). They vary in the dynamics and how they interact with the somatic compartment. (1) **AMPA synapses** are composed of a DPI block and are connected to the input of the neuron, modeling biological AMPA receptor; (2) **NMDA synapses**, also an excitatory synapses and linked to neuron's input, however, unlike AMPA synapses, it incorporates a pair of differential blocks to introduce a voltage gating mechanism. This mechanism blocks the synaptic current until the neuron's membrane potential reaching a specific threshold ( $I_{nmda\_thr}$  parameter in ARCANA); (3) **GABA<sub>A</sub>, an inhibitory synapses**, implements a current mirror to subtract presynaptic current from the input of the neuron, and (4) **GABA<sub>B</sub> synapses** has the current mirror directly connected to the somatic compartment, thus subtracting presynaptic current from membrane current, mimicking a shunting behavior.

The dynamic behavior of the DPI circuit can be approximated as a first-order differential in equation [5]:

$$\tau \frac{d}{dt} I_{syn} + I_{syn} = \frac{I_g}{I_\tau} I_w \quad (1)$$

where  $\tau = CU_T/\kappa I_\tau$  is the synapse decay time constant,  $C$  the synapse capacitance,  $U_T$  the thermal voltage,  $\kappa$  the transistor subthreshold slope factor,  $I_\tau$  the leakage current,  $I_g$  the DPI filter gain, and  $I_w$  the base weight current, and  $I_{syn}$  the synaptic current. These parameters are tunable and shared among synapses of the same type within each core.

### 2.2. The DPI neuron

We derived the equations for the neuron model by directly mapping its circuit components. The DPI neuron circuit comprises four main blocks:

- **Input DPI Model Leak:** The module integrates the DPI inputs from the synapses and the constant DC input current to charge the capacitor  $C_{mem}$ , which represents the neuron's leak conductance. It has a series of transistors that control the input current gain, amplitude of DC current, and leakage current that discharges the capacitor. The neuron's input current is composed of the constant DC current plus the synaptic currents aggregated from *AMPA* and *NMDA*, as well as *GABA<sub>A</sub>*. Additionally, the *GABA<sub>B</sub>* synapse is connected to the leakage current, directly discharging the capacitor.
- **After-HyperPolarization (AHP) block:** This block provides slow negative feedback that models spike frequency adaptation. When a postsynaptic spike occurs, it integrates the event into a recurrent negative AHP current, which is subtracted from the input, effectively suppressing the activity of the neuron.
- **Positive Feedback and Spike Generation Block:** This block mimics activation and inactivation of sodium channels through a positive feedback circuit. When the neuron

current starts to ramp up, the current driving the inverter circuits is fed back into the capacitor  $C_{mem}$ , further increasing the neuron's membrane current. Consequently, the  $I_{mem}$  current grows exponentially until the inverter circuits finish switching, at which point a spike is generated and the reset block is activated.

- **Reset Block:** This block mimics the potassium channels. After a spike event, it resets the membrane current by creating a short circuit to ground, which discharges the neuron's membrane capacitance and directs the membrane current  $I_{mem}$  to ground. This discharge period, controlled by a bias parameter, prevents  $C_{mem}$  from recharging, resulting in an absolute reset and a refractory period. Once this period ends, the  $I_{mem}$  current recharges the capacitor  $C_{mem}$  and the neuron resumes integrating its input spikes.

The dynamics of these modules are governed by the following differential equations [5]:

$$\left(1 + \frac{I_g}{I_{mem}}\right) \tau \frac{dI_{mem}}{dt} + I_{mem} \left(1 + \frac{I_{ahp}}{I_\tau}\right) = I_\infty + f(I_{mem}) \quad (2a)$$

$$\tau_{ahp} \frac{dI_{ahp}}{dt} + I_{ahp} = I_{ahp\infty} u(t) \quad (2b)$$

$$I_\infty = \frac{I_g}{I_\tau} \left( I_{in} - I_{ahp} - I_\tau \right) \quad (2c)$$

where  $I_{mem}$  denotes the subthreshold membrane current.  $I_{ahp}$  is the after-hyperpolarization current responsible of spike-frequency adaptation.  $I_\infty$  is the maximum current a neuron would reach asymptotically.  $I_\tau$  represents leakage current of the neuron,  $\tau$  is the neuron time constant, and  $I_{in}$  is the total input current from both synaptic sources and the constant DC input. All these parameters can be configured both on DYNAP-SE chip as well as the simulator.

The term  $f(I_{mem})$  in equation 2a represents the positive feedback current as described in section 2.2, and it is well modeled by an exponential function. [9]:

$$f(I_{mem}) = \frac{I_{fb}}{I_\tau} (I_{mem} - I_g) \quad I_{fb} = \frac{I_0^{\frac{1}{\kappa+1}} I_{mem}^{\frac{\kappa}{\kappa+1}}}{1 + e^{-\alpha(I_{mem} - I_{th})}} \quad (3)$$

where  $\alpha$  and  $I_{th}$  are hyper-parameters,  $I_0$  is dark current and  $\kappa$  is transistor slope factor.

*Parameters:* The digital-to-analog (DAC) bias-generator circuits configure bias parameters by generating specific magnitudes of current to govern neuronal and synaptic properties, including time constant, refractory period, and synaptic weights [6]. Internally, these parameters are represented as “coarse” and a “fine” value, where  $coarse \in [0, 7]$  and  $fine \in [0, 255]$ . These parameters exhibit variability, also known as mismatch, due to fabrication non-idealities, which inherently limits user control (see Section 3). Moreover, the assigned values are non-observable and cannot be directly retrieved from the circuits, so we rely on oscilloscope traces for measurement. The bias parameters that can be set on simulation are listed in Table 1. In the case of  $I_{th}$  and  $\alpha$  they are fitting parameters of the positive feedback and not a direct bias in DYNAP-SE.

Table 1: ARCANA model parameter sand DYNAP-SE biases correspondence.

ARCANA	DYNAP-SE bias	Description
Igain_mem	IF_THR_N	neuron gain
Itau_mem	IF_TAU1_N	Membrane time constant
refractory <sup>1</sup>	IF_RFR_N	refractory period
Idc	IF_DC_P	DC input
Itau_ampa	NPDPIE_TAU_F_P	AMPA synaptic time constant
Igain_ampa	NPDPIE_THR_F_P	AMPA synaptic gain
Iw_ampa	PS_WEIGHT_EXC_F_N	AMPA synaptic base weight current
Inmda_thr	IF_NMDA_N	NMDA voltage gating
Itau_nmda	NPDPIE_TAU_S_P	NMDA synaptic time constant
Igain_nmda	NPDPIE_THR_S_P	NMDA synaptic gain
Iw_nmda	PS_WEIGHT_EXC_S_N	NMDA synaptic base weight current
Itau_gabaa	NPDPII_TAU_F_P	GABA_A synaptic time constant
Igain_gabaa	NPDPII_THR_F_P	GABA_A synaptic gain
Iw_gabaa	PS_WEIGHT_INH_F_N	GABA_A synaptic base weight current
Itau_gabab	NPDPII_TAU_S_P	GABA_B synaptic time constant
Igain_gabab	NPDPII_THR_S_P	GABA_B synaptic gain
Iw_gabab	PS_WEIGHT_INH_S_N	GABA_B synaptic base weight current

<sup>1</sup> The refractory period in ARCANA is in milliseconds compared to the bias current in DYNAP-SE.

### 2.3. Calibration

We performed a one-time calibration (per chip) to align the simulation parameters (in pA) with those on the chip (in a.u.) using incremental adjustments. We primary calibrated the membrane leak time constant,  $I_\tau$ , because it is the only bias that directly influences the membrane current —a quantity measurable by an oscilloscope. All other parameters were kept fixed to ensure consistency during the calibration process.

Similarly, we also evaluated the above-threshold neural response by increasing the DC input ( $I_{DC}$ ) and neuron gain ( $I_{th}$ ). A slight adjustment of the leak current ( $I_\tau = 4.1\text{pA}$ ) was sufficient to match the simulation and silicon traces, reducing the neuron’s decay time constant. This calibration resulted in the values:  $I_{th} = 500\text{pA}$ ,  $I_\tau = 4.1\text{pA}$ , and  $I_{DC} = 36.6\text{pA}$  for simulation. These current values are obtained by setting a [coarse,fine] pair that controls the on-chip DAC current-bias generator [6]:  $I_{th} = [6, 88]$ ,  $I_\tau = [6, 22]$ , and  $I_{DC} = [2, 57]$ . Figure 1a shows a close match between the simulation traces (dashed orange line) and the silicon neuron traces (solid blue line), confirming that the simulation equations accurately capture the parameters and that their influence on silicon neuron behavior is faithfully modeled.

Following a similar approach, we calibrated the synapse parameters by recording the neuron’s membrane current in response to a single spike. For the AMPA synapse, the membrane current trace (Figure 1b) was obtained with  $I_\tau = 8\text{pA}$  ([coarse = 2, fine = 20]),

$I_{th} = 6.5\text{pA}$  ( $[coarse = 2, fine = 20]$ ) and  $I_w = 50\text{pA}$  ( $[coarse = 5, fine = 99]$ ). For the GABA<sub>A</sub> inhibitory synapse (Figure 1c), we set  $I_\tau = 20\text{pA}$  ( $[coarse = 2, fine = 80]$ ),  $I_{th} = 350\text{pA}$  ( $[coarse = 4, fine = 80]$ ),  $I_{DC} = 6.5\text{pA}$  ( $[coarse = 2, fine = 10]$ ) and  $I_w = 400\text{pA}$  ( $[coarse = 5, fine = 200]$ ). Note that for inhibitory synapse, the neuron's membrane current must be above its resting level to observe the inhibitory effect.

We observed a higher deviation in traces for AMPA ( $MSE = 1.888 \times 10^{-7}$ ) and GABA<sub>A</sub> synapses ( $MSE = 1.879 \times 10^{-6}$ ) compared to the traces with constant input current. This discrepancy may have arisen from the scale of the simulated voltage. When the input is provided via a synapse, the somatic voltage only reaches about 80 mV, whereas with constant DC current, it can rise to 800 mV due to a positive feedback mechanism. Another possible explanation is that parasitic capacitance from the circuit layout and fabrication, which is not accounted for in the simulation, could contribute to the difference.

Here, we calibrated a single neuron such that all others remain within a 20% margin of variability caused by mismatch. A more rigorous approach would involve a statistical analysis of the entire population, as described in [28] using chip-in-loop calibration, but this is beyond the scope of the present work.

This calibration process is integrated in subsequent experiments (see sections: 4.2 and 4.3). These experiments serve as validation tests, where a network trained in simulation was deployed on the chip and produced outcomes similar to those observed in ARCANA.

### 3. Hardware mismatch

Inherent variability in the analog substrate can cause computational discrepancies relative to idealized simulations, imposing constraints that are beyond the user's control. Typically, the coefficient of variation for all parameters ranges from 10 – 20% [28].

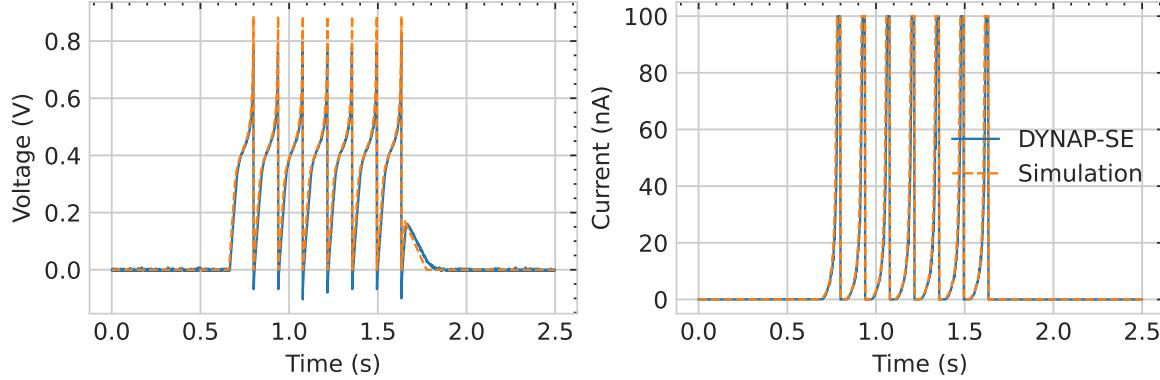
This mismatch is evident while observing from membrane leak. Figure 2 shows the distribution of time constant ( $\tau_{mem}$ ) computed from membrane traces with bias settings applied via `IF_TAU1_N`. The coarse bias is fixed at 5, while the fine bias is swept from 74 to 190. Notably, as the  $\tau$  value is modified, its standard deviation decreases proportionally.

ARCANA incorporates mismatch in both neuron and synapse models, allowing users to assess their models' resilience to parameter variability. Users can select the desired degree of variability, starting with deterministic simulations for initial testing and gradually introducing mismatch as needed.

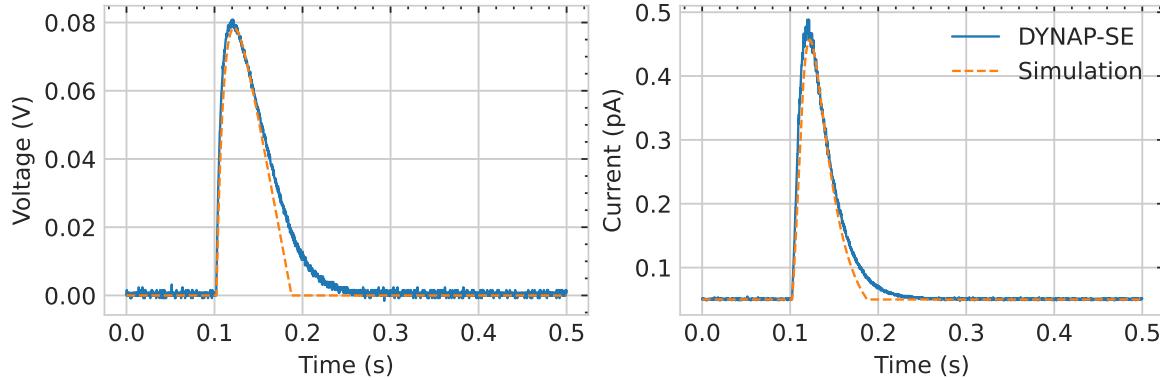
### 4. Results

We present three experiments that assess the dynamics and accuracy of our PyTorch-based implementation of the DPI neuron and synapse model. The first experiment, described in Section 4.1, features a frequency resonator used solely to validate the training process in simulation; it was not deployed on the chip. In all experiments, we employed the *autograd* tool to automatically calculate gradients, thereby facilitating the tuning and optimization of complex models.

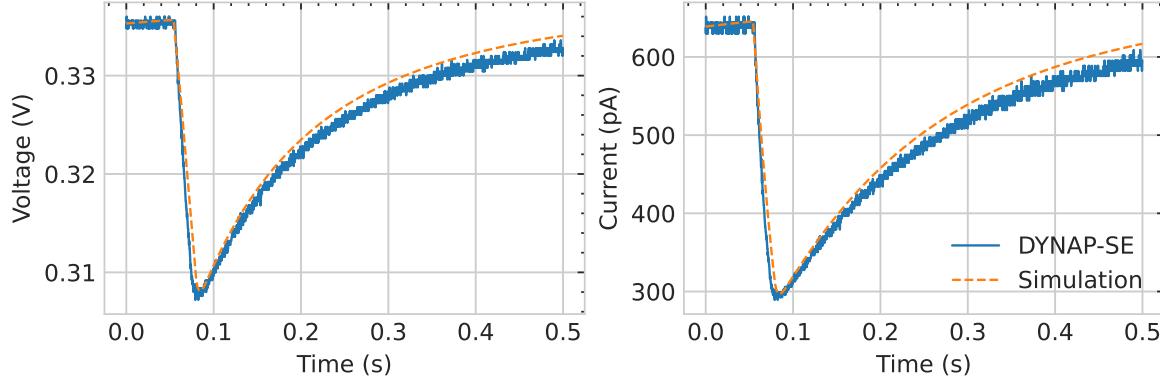
Figure 1: ARCANA and DYNAP-SE comparison, where on each figure, the *left pane* represents the voltage and the *right pane* the soma current.



(a) Using a constant DC input, we compared ARCANA simulator with the DYNAP-SE chip . For visualization purposes, the current was capped at  $100\text{ nA}$ , although in practice it would rise exponentially until the firing threshold was reached.



(b) Comparison between ARCANA and DYNAP-SE chip using a AMPA input synapse. The traces exhibit a mean squared error ( $MSE$ ) of  $1.888 \times 10^{-7}$ .



(c) Comparison between ARCANA and DYNAP-SE chip using a GABA<sub>a</sub> input synapse. The traces exhibit a mean squared error ( $MSE$ ) of  $1.879 \times 10^{-6}$ .

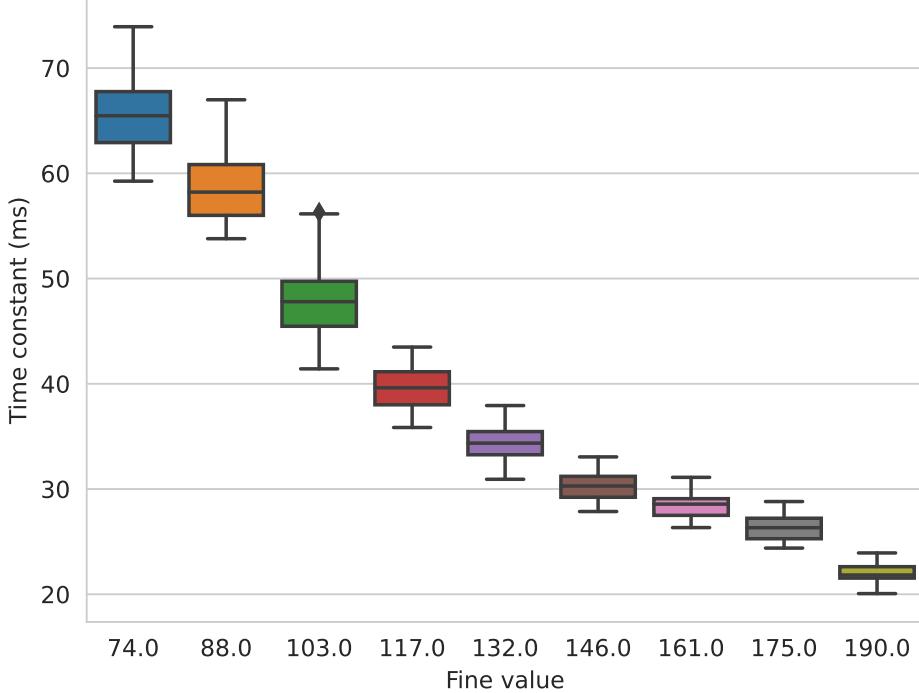


Figure 2: Time constant distribution for  $I_r$  ranging from 530 nA to 1676 nA.

Once we validated the training process, we trained a small network to perform binary classification using Backpropagation Through Time (BPTT). The trained network was then deployed on DYNAP-SE chip (see section: 4.2). In the last experiment (section: 4.3), we present a proof-of-concept for Event-based Three-factor Local Plasticity (ETLP) [20]. In this experiment, a network was trained in simulation using ETLP to distinguish between two spike patterns. After training, we deployed it on the DYNAP-SE chip for inference.

These experiments serve as validation for the algorithm’s compatibility with hardware-aware training. Deploying pre-learned weights and synaptic connections on mixed-signal hardware paves the way for innovative real-world applications in neuromorphic computing. Notably, before deployment, it is critical to perform a one-time manual calibration per chip, as detailed in Section 2. In future work, this calibration process can be automated using autograd and a chip-in-loop paradigm [13].

#### 4.1. Spike frequency resonator

In this experiment, we perform a basic task to showcase the use of the *autograd* tool to optimize neuron parameters. The task involves constant current injection to a neuron with  $I_{DC} = 10\text{pA}$ . The objective is to modify neuron gain ( $I_{th}$ ) and membrane time constant ( $I_\tau$ ) to achieve the desired firing frequency of 2.5 Hz. Figure 3 (top) shows the initial response of the neuron to DC injection (solid blue line). The absence of a spike response indicates the need to adjust the neuron’s parameters. Instead of manually calibrating these parameters, we use autograd to automatically determine the optimal values so that the neuron fires at the user-specified

Table 2: Parameters used for spike frequency resonator experiment.

Parameter	Value
Initial neuron $I_\tau$	$4e^{-12}$
Initial neuron $I_g$	$20e^{-12}$
$I_{DC}$	$10e^{-12}$
Positive feedback gain ( $I_g$ )	$20e^{-12}$
Positive feedback normalization ( $\alpha$ )	$2.0e^9$
Learning rate	$5e^{-3}$
Optimizer	Adam

frequency of 2.5 HZ.

The loss function is the mean squared error between the number of output spikes and the target frequency. This loss is calculated with respect to the neuron parameters that we aim to optimize—neuron gain ( $I_{th}$ ) and membrane time constant ( $I_\tau$ ). Figure 3 (bottom) illustrates that the training was completed (loss  $\simeq 0$ ) in fewer than 40 epochs. Figure 3 (top) also shows neuron’s response post-optimization in orange.

#### 4.2. Binary Image Classification

In this experiment, we trained the network to perform image classification for digit recognition. For simplicity, we only consider 0 and 1 of the MNIST dataset. We trained the connection matrix of the network and subsequently deployed the trained network on DYNAP-SE for on-chip inference. The aim is to ensure that the designated neuron on the chip exhibits a higher firing rate than the baseline for its preferred digit (0 or 1). The network receives the pixel intensity encoded as firing rate with a Poisson distribution as shown in Figure 4.

**Syncing ARCANA and DYNAP-SE:** To ensure a faithful deployment of a trained network, it is essential to match the leakage and gain currents for both neurons and synapses on the chip and the ARCANA simulator, as described in section 2.3. On the DYNAP-SE chip, the base weight for each synapse type can be tuned using bias currents, similar to other neuron and synaptic parameters (see section: 2.2). This base weight is common to all neurons in a single core. Hence, to modify the weights for a specific neuron, we alter the number of synapses connecting two neurons (connection matrix). This necessitates the quantization of the learned connection matrix into integer values. We dealt with quantization constraint by employing Quantization-Aware Training (QAT) procedure.

**Network training:** The classification task discussed here is linearly separable, therefore, we train a single-layer network with 256 input channels and 2 readout neurons, with one neuron representing digit 0 and the other representing digit 1. Each neuron can receive input with multiple excitatory (AMPA) and inhibitory (GABA<sub>A</sub>) synapses. We initialized AMPA and

# Neuron parameters optimization

Optimizing leakage and gain currents of the neuron in order to fire 5 spikes in 2 seconds with a DC input of  $10\text{pA}$

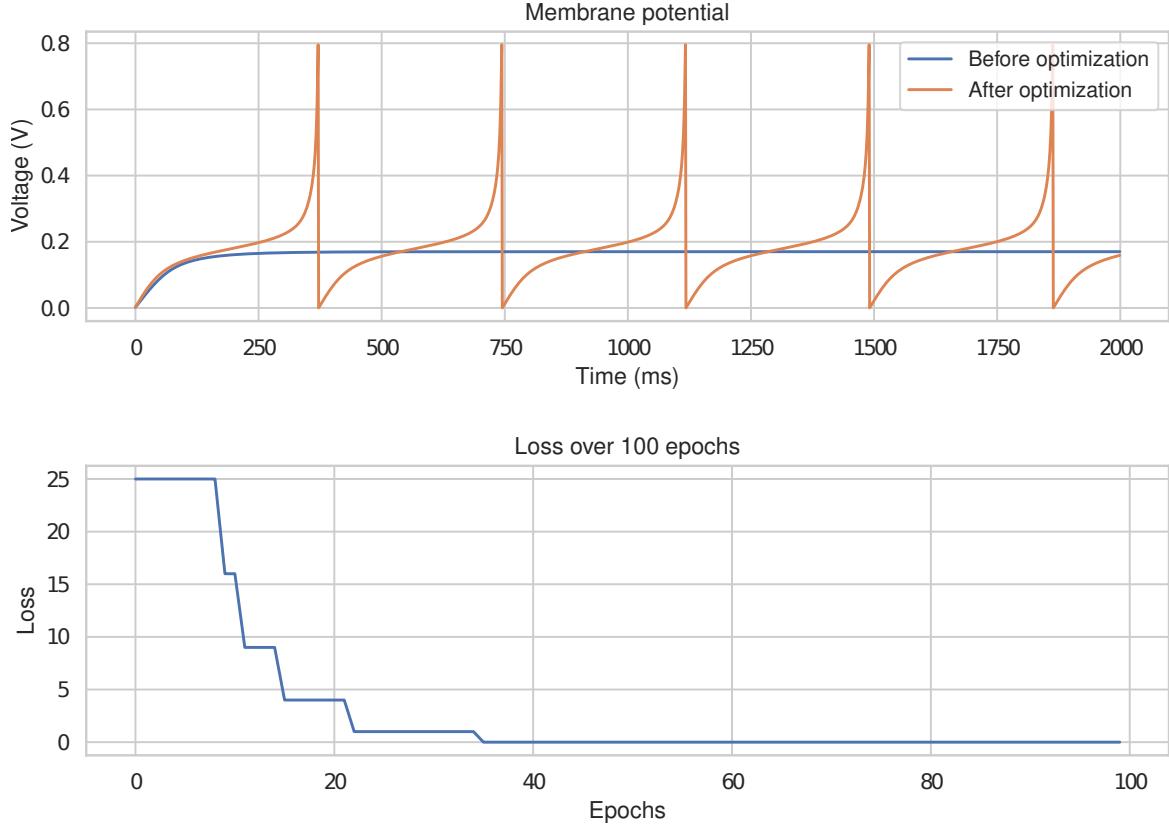


Figure 3: Neuron parameters optimization to obtain an output frequency of  $2.5\text{Hz}$  from a constant input current of  $10\text{pA}$ .

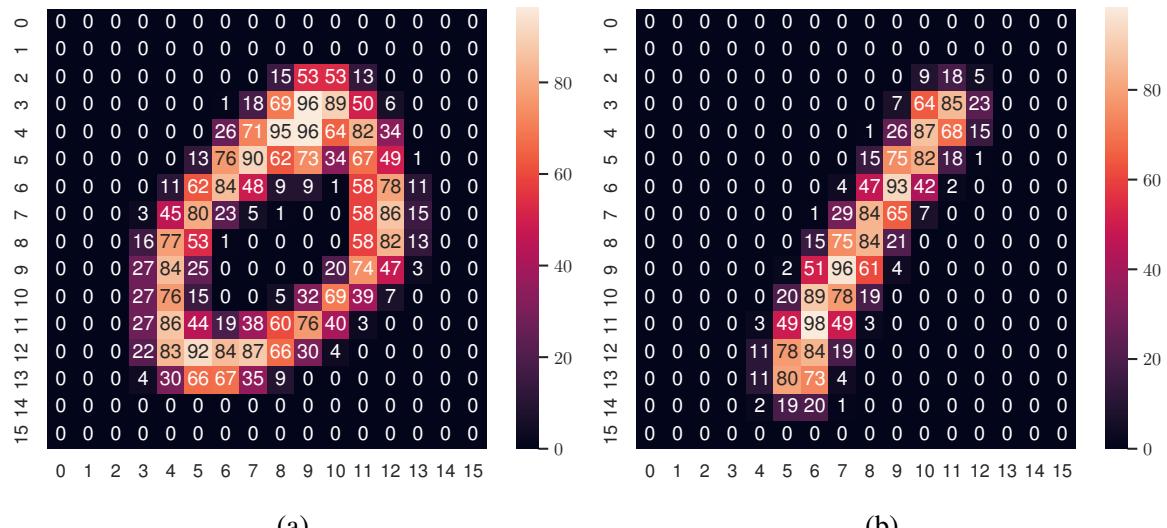


Figure 4: Samples from class 0 (A) and 1 (B) with the corresponding frequency (Hz) value to convert it into a spike train.

GABA<sub>A</sub> connection matrices with a uniform distribution. We pruned the synapses to meet the fan-in limitations (64 per neuron) of the chip. We introduced mismatch in the synaptic weights bias ( $I_w$ ) during training. However, all other neuron and synaptic parameters were manually calibrated (3) and remain unchanged during training. Note that the framework also supports incorporating mismatch across all parameters during training, even though this feature is not utilized in this experiment.

Training with 32-bit floating-point weights provides a broader dynamic range but complicates on-chip deployment, as DYNAP-SE uses quantized weights. With the help of quantization techniques such as QAT [30], we reduced the precision of weights from float to integer, resulting in efficient computation while maintaining high accuracy during inference on the chip. A rounding operator was used to map the floating point tensor to a quantized representation:

$$x_q = \text{round}(x_f)$$

During training, QAT introduces a “mock” low precision in the forward pass, while the backward pass remains full precision (Figure 5a). To deal with the quantization operation gradient, the Straight-Through Estimator (STE) surrogate gradient is used. This approach allows the gradient to be transmitted unaltered through the fake-quantization operator.

As previously stated, the DYNAP-SE chip also restricts the fan-in of a neuron, permitting only 64 input synapses per neuron. This restriction can differ for different chips, however it must be taken in account during training to avoid exceeding the synapse limit and risking a topology that is incompatible with the hardware. To address this challenge, we penalize the model for higher fan-in by adding L1 and/or L2 regularization terms. For instance, for  $W$  as the connection matrix and  $\text{fan\_in}$  as the desired number of input synapses (here  $\text{fan\_in} = 64$ ), we used a L1 regularization, where  $\text{L1} = \lambda \sum |W - \text{fan\_in}|$ .  $\lambda$  is a hyperparameter that controls the strength of the regularization. If the sum of the rounded weights is not yet exactly equal to the desired value, a final adjustment can be made by adding a constant value  $C$  to each weight, defined as  $C = (\text{fan\_in} - \sum(\text{round}(W))) / N$ . Here  $N$  is the number of input neurons, and the sum is over all the input weights of each postsynaptic neuron.

**Loss:** The loss function for this task is a softmax cross-entropy applied to the neuron inputs. Here, the network is optimized to increase the input current of the output neuron corresponding to the presented class while reducing the current of the other neuron. We defined the loss function on the total input current of each neuron (difference between the AMPA and GABA<sub>A</sub> current) in Equation 4a.

$$L = - \sum_{n=1}^N \sum_{q=1}^Q y_{n,q} \log \frac{e^{x_{n,q}}}{\sum_{i=1}^Q e^{x_{n,i}}} \quad (4a)$$

$$x_n = \sum_{t=0}^T I_{\text{AMPA}}(t)_n - I_{\text{GABA}_A}(t)_n \quad (4b)$$

$$(4c)$$

Table 3: Parameters used for binary classification problem.

Parameter	Value
Neuron $I_\tau$	$1.8e^{-12}$
Neuron $I_g$	$45e^{-12}$
$I_{DC}$	$240e^{-12}$
Positive feedback gain ( $I_g$ )	$225e^{-12}$
Positive feedback normalization ( $\alpha$ )	$1.0e^9$
GABA <sub>a</sub> /AMPA synapse $I_\tau$	$4e^{-12}$
GABA <sub>a</sub> /AMPA synapse $I_g$	$10e^{-12}$
GABA <sub>a</sub> /AMPA synapse base weight	$400e^{-12}$
Learning rate	$1e^{-3}$
Optimizer	SGD
Maximum number of input synapses	40

Where  $y_{n,q}$  is the target class probability for sample  $n$  and class  $q$ ,  $x_n$  is the difference between the accumulation over time of the AMPA and GABA<sub>a</sub> currents over time,  $Q$  the number of classes,  $N$  the number of samples, and  $T$  the total simulation time.

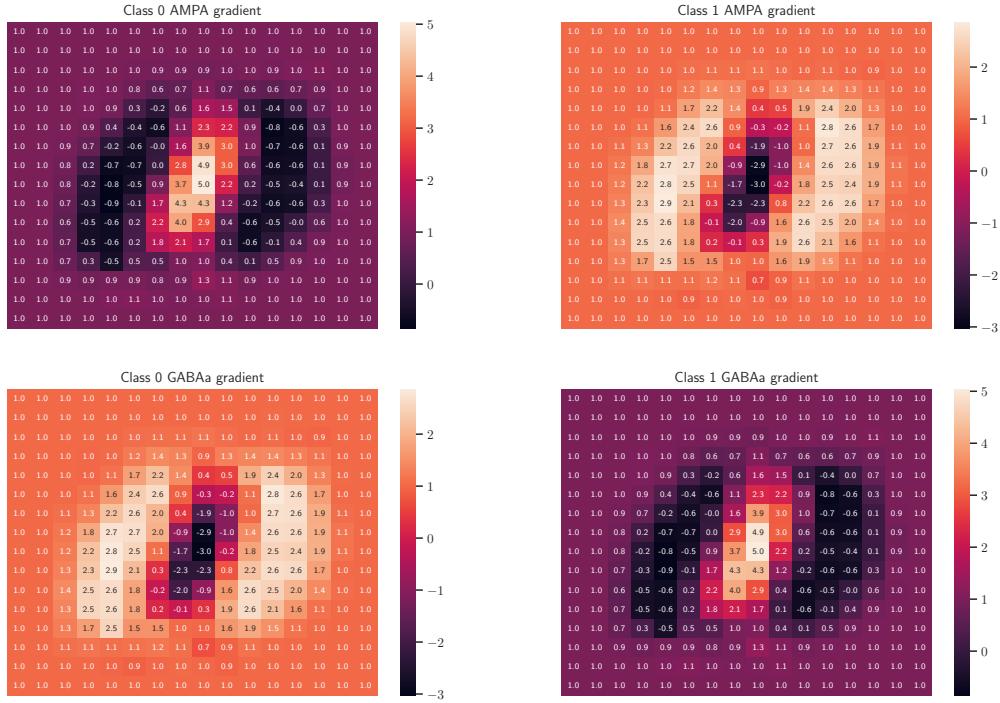
The final connection matrix obtained after training is then rounded, as shown in Figure 5b. Before deploying the network on the DYNAP-SE, we performed an inference mode evaluation in simulation (Figure 6a). As anticipated, the neurons spike behavior only when exposed to their respective selective digits.

The network was deployed on the hardware as a last step in this process. During on-chip inference, 2115 test samples of MNIST digits (0 or 1) were randomly selected and presented to the network for 50 ms each, interleaved with 50 ms rest periods with no stimulus. The firing response of the neurons was recorded for each presentation. Figure 6b shows representative traces recorded on the chip during one inference session. Running entirely on hardware, the emulated network achieved an accuracy of 99.11% in predicting the stimulus. This demonstrates the feasibility of training network connections in software and efficiently deploying it on a mixed-signal chip.

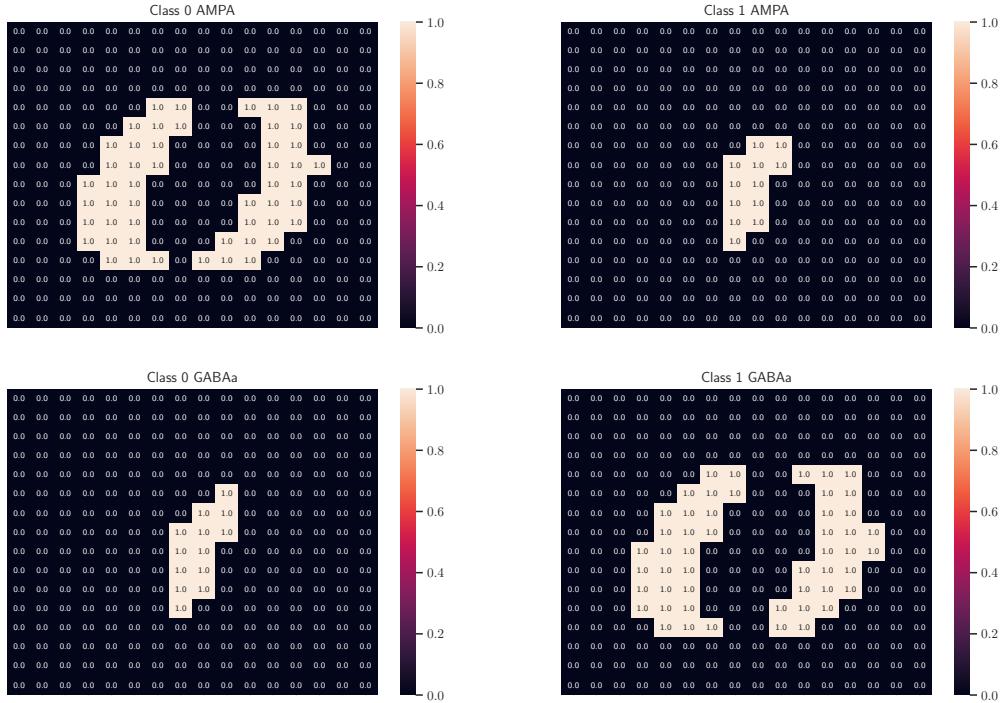
#### 4.3. Learning rules for DYNAP-SE

In this experiment, we showed how ARCANA is a promising tool-chain for end-to-end network training and on-chip inference post-deployment. An important advantage of having a simulator for a mixed-signal processor such as DYNAP-SE is the possibility of testing various learning rules such as ETLP [20] with the neuron and synapse model implemented on the hardware. Thus, it serves as a proof-of-concept to test the viability of these learning algorithms for specific hardware before designing the circuit layout of the learning rule for on-chip implementation.

In this experiment, we trained a network using ETLP [20] and deployed the final weights on DYNAP-SE. ETLP is a supervised three-factor local learning rule composed of 1) a presynaptic

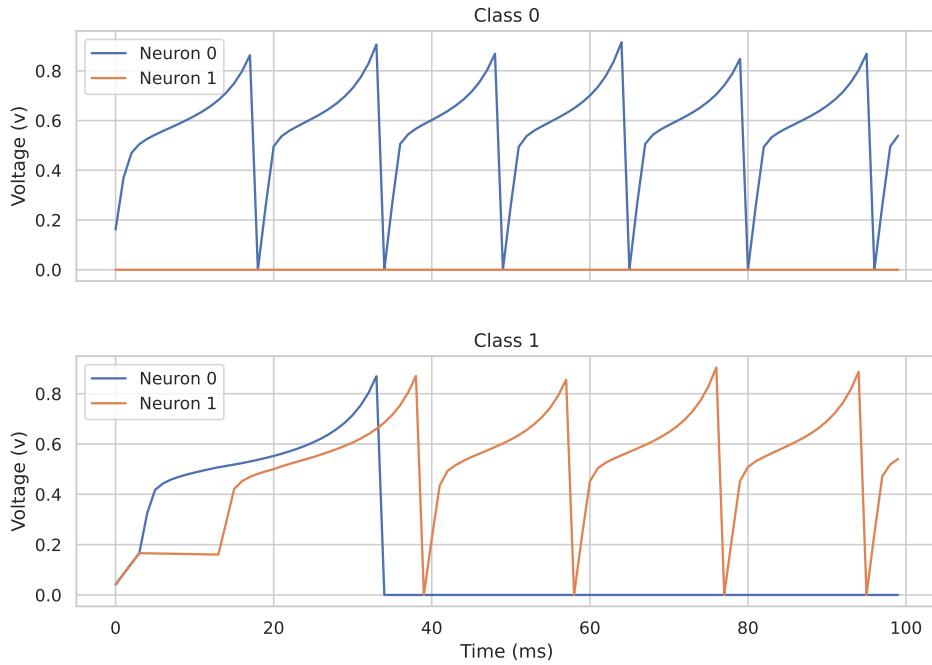


(a) AMPA and GABA<sub>a</sub> synapses gradients during the training process for classes 0 and 1. The value of the gradients are stored and calculated in full-precision.

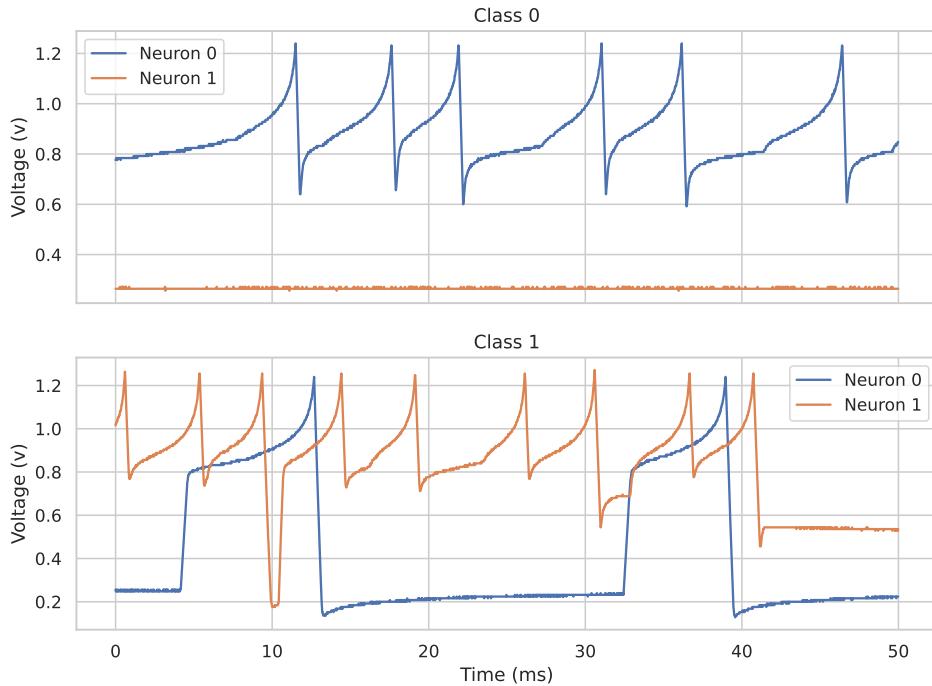


(b) AMPA and GABA<sub>a</sub> final connection matrix for class 0 and 1 obtained after training and quantizing them.

Figure 5: AMPA and GABA<sub>a</sub> gradients and final quantized weight matrices.



(a) Output neurons voltages simulated in ARCANA when receiving a sample from class 0 and class 1.



(b) Output neurons voltages recorded on DYNAP-SE when receiving a sample from class 0 and class 1. The figure shows the output of the neuron voltage recorded directly from the hardware, from a sample of each class.

Figure 6: Output neuron voltages for classes 0 and 1 recorded in simulation (a) and on chip (b).

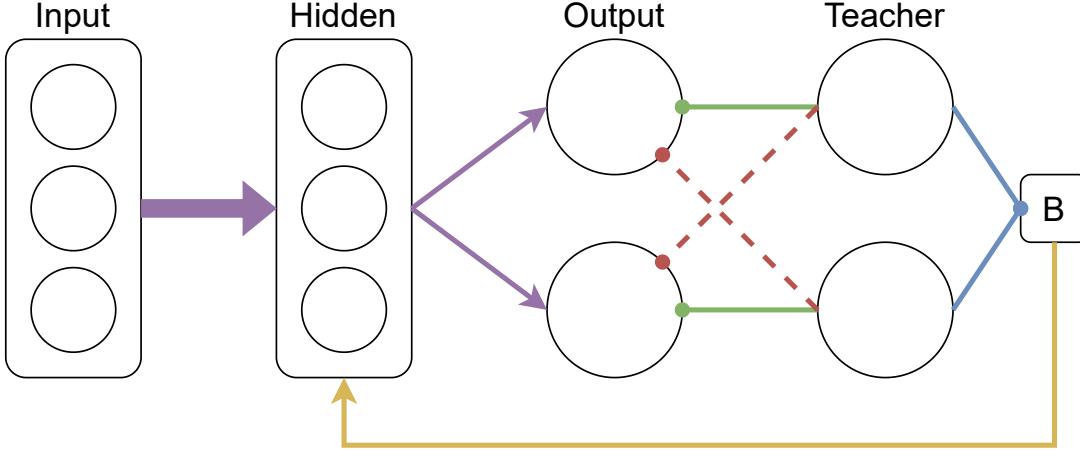


Figure 7: Architecture of ETLP learning rule. The teaching neurons are connected to the output neurons via excitatory (green) and inhibitory synapses (dashed red) to learn the weights between the hidden and output layers. The teaching neurons are also connected through a random initialized weight matrix B to the hidden layer neurons to apply the learning signal to the weight matrix between the input and the hidden layer.

trace, 2) a postsynaptic value, and 3) an external teaching signal from teacher neurons that triggers the learning (Figure 7). The weight update follows equation 5b.

$$\tau_\epsilon \frac{d\epsilon}{dt} = \epsilon + x \quad (5a)$$

$$\Delta W = \sigma'(v(t))\epsilon(t)I(t) \quad (5b)$$

where  $\Delta W$  is the weight change,  $\epsilon$  is the pre-synaptic trace computed by accumulating the input spikes,  $\sigma'(v) = (1 + \alpha|v|)^{-2}$  is a non-linear function depending of the soma voltage  $v(t)$  [29],  $x$  the input pre-synaptic spike,  $\tau_\epsilon$  the pre-synaptic trace time constant and  $I(t)$  the input from the teaching neurons.

In this experiment, we trained a network (as illustrated in 7) to distinguish between two input frequencies —1 Hz and 10 Hz per channel. This system can be applied to anomaly detection. The network architecture consists of 50 input neurons, 50 hidden neurons, and 50 output neurons, along with 2 teacher neurons. A high-frequency teacher signal at 50 Hz facilitates learning of the preferred class, while a low-frequency teacher at 5 Hz simulates noise. Synaptic weight changes follow the ETLP rule. For inference, the trained weights were deployed on the DYNAP-SE chip.

Figure 8 shows spike activity of the whole network during the training and inference process in simulation. The output neurons fire independently for the initial 4 seconds of the pattern presentation. In the next 20 seconds, both tasks are presented alternatively in the network by increasing the firing rate of the input neurons from each class. This allows the hidden and output layers to learn the relation between the classes and the input pattern. During

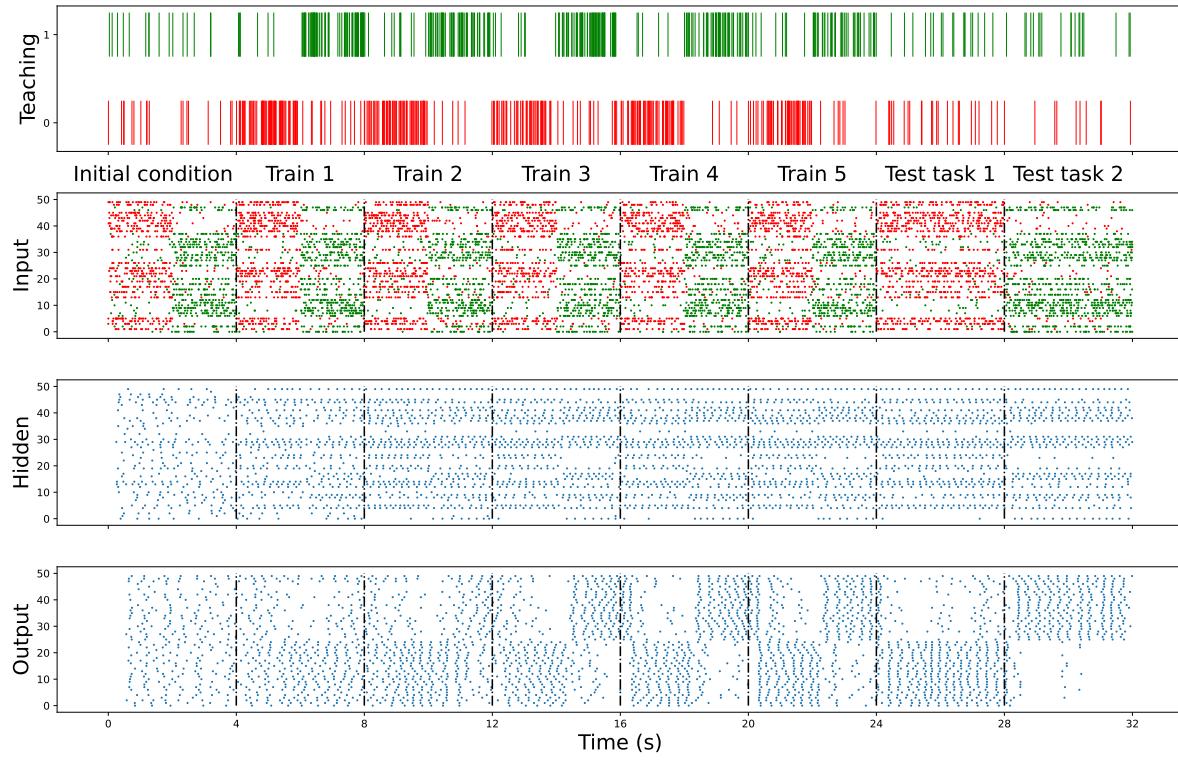


Figure 8: Raster plot of the simulation performed on ARCANA to train a network using ETLP. The first plot represents the spikes of the teaching neurons that are connected to the hidden and output layers. The second plot corresponds to the input pattern, which has two neuron groups that fire with different frequencies. The network has to learn to associate each class with each neuron group. The third and fourth plots are the spikes produced by the hidden and output neurons during the training and testing processes.

Table 4: Parameters used for ETLP experiment. The presynaptic trace time constant is the same as the neuron time constant.

Parameter	Value
Neuron $I_\tau$	$4.32e^{-12}$
Neuron $I_g$	$20e^{-12}$
$I_{DC}$	0
Positive feedback gain ( $I_g$ )	$225e^{-12}$
Positive feedback normalization ( $\alpha$ )	$1.0e^9$
GABA <sub>a</sub> /AMPA synapse $I_\tau$	$4e^{-12}$
GABA <sub>a</sub> /AMPA synapse $I_g$	$10e^{-12}$
GABA <sub>a</sub> /AMPA synapse base weight	$200e^{-12}$
Maximum number of input synapses	40

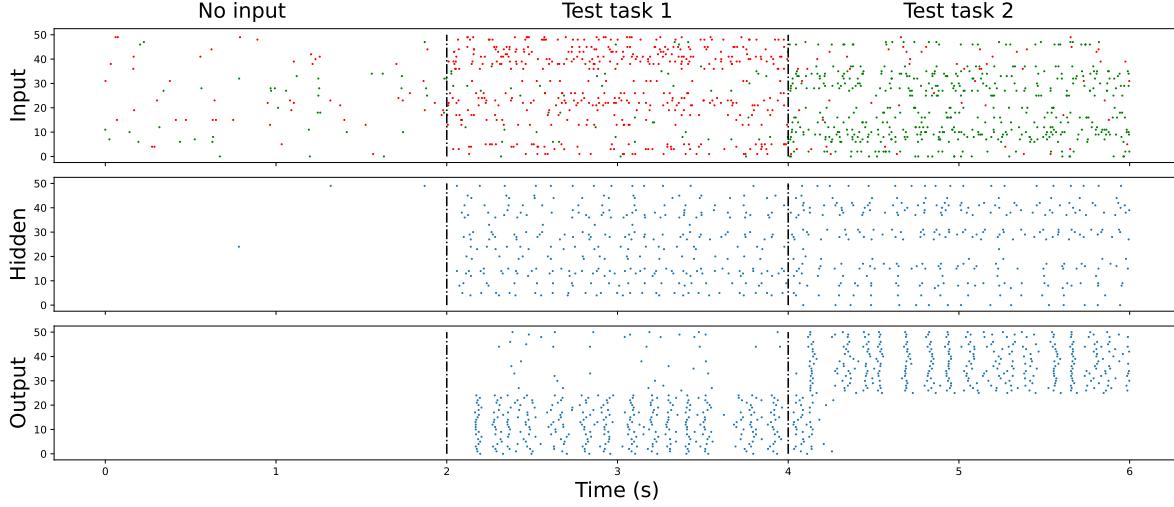


Figure 9: Raster plot of the network running on DYNAP-SE chip with the weights obtained by training in simulation using ETLP. The network was first trained on simulation and then deployed to the chip.

inference, the teacher neurons are set to a low-rate firing state (5 Hz), a subset of neurons in output layer exhibit higher firing response for their preferred class, validating correct association was learned during training.

**On-chip inference:** Once the network was trained and validated for inference in simulation, the trained weights were mapped into a network created on DYNAP-SE chip without any preprocessing step. Figure 9 shows the spike activity of the whole network on chip, exhibiting the desired behavior for each individual input. During on-chip inference, with no teacher provided, the bottom raster plot shows two distinct neuron groups, each exhibiting increased activity when their preferred pattern is presented.

## 5. Discussion

In this paper, we presented the ARCANA simulator for mixed-signal hardware and validated its performance using the DYNAP-SE chip. The advantage of ARCANA is its ability to use of the PyTorch “autograd” feature, which allows the optimization of the internal parameters of the network. While other similar simulators developed in parallel to this work have been proposed within the Rockpool framework [3], specifically tailored for DYNAP-SE2 chip [22], ARCANA boasts a more versatile application and is generic to any processor incorporating DPI based neuron models, independent of the specific simulation framework used. This adaptability of ARCANA is achieved by fine-tuning parameters such as the positive-feedback exponential function parameters and the chip constant values, including the transistor slope factor and neuron capacitor. Furthermore, ARCANA allows us to optimize not only the weights but other parameters as well. In addition, ARCANA includes the possibility to apply mismatch into the neuron and synapse parameters found in the hardware. This mismatch can be included in the

training framework, making the system more robust to variability in the biases. It also provides a more realistic environment for simulation, where the non-idealities of the hardware are taken into account. Finally, we demonstrated the feasibility of implementing ETLP in hardware and applied it to online training in mixed signal hardware. This opens new doors to the development of new local learning rules in embedded systems that go beyond the ones presented so far [10].

## Acknowledgements

F.M.Q. was supported by FPU grant (FPU18/04321) from the Spanish Ministry of Universities. This work is supported by the HORIZON EUROPE EIC Pathfinder Grant ELEGANCE (Grant No. 101161114), and has received funding from Swiss National Science Foundation (SNSF 200021E\_222393). This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

## Competing interests

The authors declare no competing interests.

## Code availability

<https://github.com/ferqui/ARCANA>

## References

- [1] C. Bartolozzi and G. Indiveri. “Synaptic dynamics in analog VLSI”. In: *Neural Computation* 19.10 (Oct. 2007), pp. 2581–2603. doi: 10.1162/neco.2007.19.10.2581.
- [2] Romain Brette and Wulfram Gerstner. “Adaptive exponential integrate-and-fire model as an effective description of neuronal activity”. In: *Journal of neurophysiology* 94.5 (2005), pp. 3637–3642. doi: 10.1152/jn.00686.2005.
- [3] Ugurcan Cakal, Maryada, Chenxi Wu, Ilkay Ulusoy, and Dylan Richard Muir. “Gradient-descent hardware-aware training and deployment for mixed-signal neuromorphic processors”. In: *Neuromorphic Computing and Engineering* 4.1 (Mar. 2024), p. 014011. doi: 10.1088/2634-4386/ad2ec3. URL: <https://dx.doi.org/10.1088/2634-4386/ad2ec3>.
- [4] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri. “Neuromorphic electronic circuits for building autonomous cognitive systems”. In: *Proceedings of the IEEE* 102.9 (Sept. 2014), pp. 1367–1388. ISSN: 0018-9219. doi: 10.1109/JPROC.2014.2313954.
- [5] Elisabetta Chicca, Fabio Stefanini, Chiara Bartolozzi, and Giacomo Indiveri. “Neuromorphic electronic circuits for building autonomous cognitive systems”. In: *Proceedings of the IEEE* 102.9 (2014), pp. 1367–1388. ISSN: 0018-9219. doi: 10.1109/JPROC.2014.2313954.

- [6] T. Delbrück and A. Van Schaik. “Bias Current Generators with Wide Dynamic Range”. In: *Analog Integrated Circuits and Signal Processing* 43.3 (2005), pp. 247–268.
- [7] Marc-Oliver Gewaltig and Markus Diesmann. “NEST (NEural Simulation Tool)”. In: *Scholarpedia* 2.4 (2007), p. 1430.
- [8] M.L. Hines and N.T. Carnevale. “The NEURON simulation environment”. In: *Neural Computation* 9.6 (1997), pp. 1179–1209.
- [9] G. Indiveri, F. Stefanini, and E. Chicca. “Spike-based learning with a generalized integrate and fire silicon neuron”. In: *International Symposium on Circuits and Systems (ISCAS)*. IEEE. Paris, France, 2010, pp. 1951–1954. doi: 10.1109/ISCAS.2010.5536980.
- [10] Lyes Khacef, Philipp Klein, Matteo Cartiglia, Arianna Rubino, Giacomo Indiveri, and Elisabetta Chicca. “Spike-based local synaptic plasticity: a survey of computational models and neuromorphic circuits”. In: *Neuromorphic Computing and Engineering* 3.4 (Nov. 2023), p. 042001. issn: 2634-4386. doi: 10.1088/2634-4386/ad05da. URL: <http://dx.doi.org/10.1088/2634-4386/ad05da>.
- [11] Corey Lammie and Mostafa Rahimi Azghadi. “MemTorch: A Simulation Framework for Deep Memristive Cross-Bar Architectures”. In: *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2020, pp. 1–5. doi: 10.1109/ISCAS45731.2020.9180810.
- [12] Ana Lebanov, Mauricio Velazquez Lopez, Florian De Roose, Nikolas P Papadopoulos, Giacomo Indiveri, Arianna Rubino, Melika Payvand, Steve Smout, Myriam Willegems, Francky Catthoor, et al. “Flexible Unipolar IGZO Transistor-Based Integrate and Fire Neurons for Spiking Neuromorphic Applications”. In: *Biomedical Circuits and Systems, IEEE Transactions on* (2023). doi: <https://doi.org/10.1109/TBCAS.2023.3321506>.
- [13] Maryada, Saray Soldado-Magraner, Martino Sorbaro, Rodrigo Laje, Dean V. Buonomano, and Giacomo Indiveri. “Stable recurrent dynamics in heterogeneous neuromorphic computing systems using excitatory and inhibitory plasticity”. In: (Aug. 2023). doi: 10.1101/2023.08.14.553298. URL: <http://dx.doi.org/10.1101/2023.08.14.553298>.
- [14] Carver Mead. “Neuromorphic Engineering: In Memory of Misha Mahowald”. In: *Neural Computation* 35 (2023), pp. 343–383. doi: 10.1162/neco\_a\_01553.
- [15] Mohammad Javad Mirshojaeian Hosseini, Yi Yang, Aidan J Prendergast, Elisa Donati, Miad Faezipour, Giacomo Indiveri, and Robert A Nawrocki. “An organic synaptic circuit: toward flexible and biocompatible organic neuromorphic processing”. In: *Neuromorphic Computing and Engineering* 2.3 (Sept. 2022), p. 034009. issn: 2634-4386. doi: 10.1088/2634-4386/ac830c. URL: <http://dx.doi.org/10.1088/2634-4386/ac830c>.
- [16] S. Moradi, N. Qiao, F. Stefanini, and G. Indiveri. “A Scalable Multicore Architecture With Heterogeneous Memory Structures for Dynamic Neuromorphic Asynchronous Processors (DYNAPs)”. In: *IEEE Transactions on Biomedical Circuits and Systems* 12.1 (Feb. 2018), pp. 106–122. doi: 10.1109/TBCAS.2017.2759700.

- [17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- [18] Christian Pehle, Sebastian Billaudelle, Benjamin Cramer, Jakob Kaiser, Korbinian Schreiber, Yannik Stradmann, Johannes Weis, Aron Leibfried, Eric Müller, and Johannes Schemmel. “The BrainScaleS-2 accelerated neuromorphic system with hybrid plasticity”. In: *Frontiers in Neuroscience* 16 (2022), p. 795876.
- [19] Jiale Quan, Zhen Liu, Bo Li, and Jiajun Luo. “Ultra-Low-Power Compact Neuron Circuit with Tunable Spiking Frequency and High Robustness in 22 nm FDSOI”. In: *Electronics* 12.12 (2023). ISSN: 2079-9292. doi: 10.3390/electronics12122648. URL: <http://dx.doi.org/10.3390/electronics12122648>.
- [20] Fernando M Quintana, Fernando Perez-Peña, Pedro L Galindo, Emre O Neftci, Elisabetta Chicca, and Lyes Khacef. “ETLP: event-based three-factor local plasticity for online learning with neuromorphic hardware”. In: *Neuromorphic Computing and Engineering* 4 (3 Aug. 2024), p. 034006. ISSN: 2634-4386. doi: 10.1088/2634-4386/AD6733. URL: <https://iopscience.iop.org/article/10.1088/2634-4386/ad6733%20https://iopscience.iop.org/article/10.1088/2634-4386/ad6733/meta>.
- [21] Malte J. Rasch, Diego Moreda, Tayfun Gokmen, Manuel Le Gallo, Fabio Carta, Cindy Goldberg, Kaoutar El Maghraoui, Abu Sebastian, and Vijay Narayanan. “A Flexible and Fast PyTorch Toolkit for Simulating Training and Inference on Analog Crossbar Arrays”. In: *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. 2021, pp. 1–4. doi: 10.1109/AICAS51828.2021.9458494.
- [22] Ole Richter, Chenxi Wu, Adrian M Whatley, German Köstinger, Carsten Nielsen, Ning Qiao, and Giacomo Indiveri. “DYNAP-SE2: a scalable multi-core dynamic neuromorphic asynchronous spiking neural network processor”. In: *Neuromorphic Computing and Engineering* 4.1 (Jan. 2024), p. 014003. doi: 10.1088/2634-4386/ad1cd7. URL: <https://doi.org/10.1088/2634-4386/ad1cd7>.
- [23] Arianna Rubino, Can Livanelioglu, Ning Qiao, Melika Payvand, and Giacomo Indiveri. “Ultra-Low-Power FDSOI Neural Circuits for Extreme-Edge Neuromorphic Intelligence”. In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 68.1 (2020), pp. 45–56. doi: 10.1109/TCSI.2020.3035575.
- [24] Philipp Spilger, Eric Müller, Arne Emmel, Aron Leibfried, Christian Mauch, Christian Pehle, Johannes Weis, Oliver Breitwieser, Sebastian Billaudelle, Sebastian Schmitt, et al. “hxtorch: PyTorch for BrainScaleS-2: perceptrons on analog neuromorphic hardware”. In: *IoT Streams for Data-Driven Predictive Maintenance and IoT, Edge, and Mobile for Embedded Machine Learning: Second International Workshop, IoT Streams 2020*,

- and First International Workshop, ITEM 2020, Co-located with ECML/PKDD 2020, Ghent, Belgium, September 14-18, 2020, Revised Selected Papers 2.* Springer. 2020, pp. 189–200.
- [25] Marcel Stimberg, Romain Brette, and Dan F.M. Goodman. “Brian 2, an intuitive and efficient neural simulator”. In: *eLife* 8 (Aug. 2019). issn: 2050084X. doi: 10.7554/eLife.47314.
  - [26] Srikanth Vuppunuthala and Vijay Shankar Pasupureddi. “3.6-pJ/Spike, 30-Hz Silicon Neuron Circuit in 0.5-V, 65 nm CMOS for Spiking Neural Networks”. In: *IEEE Transactions on Circuits and Systems II: Express Briefs* (2023). doi: 10.1109/TCSII.2023.3324584.
  - [27] Zhenming Yu, Stephan Menzel, John Paul Strachan, and Emre Neftci. “Integration of physics-derived memristor models with machine learning frameworks”. In: *arXiv preprint arXiv:2403.06746* (2024).
  - [28] Dmitrii Zendrikov, Sergio Solinas, and Giacomo Indiveri. “Brain-inspired methods for achieving robust computation in heterogeneous mixed-signal neuromorphic processing systems”. In: *Neuromorphic Computing and Engineering* 3.3 (July 2023), p. 034002. issn: 2634-4386. doi: 10.1088/2634-4386/ace64c. URL: <http://dx.doi.org/10.1088/2634-4386/ace64c>.
  - [29] Friedemann Zenke and Surya Ganguli. “Superspike: Supervised learning in multilayer spiking neural networks”. In: *Neural computation* 30.6 (2018), pp. 1514–1541.
  - [30] Neta Zmora, Hao Wu, and Jay Rodge. *Achieving FP32 Accuracy for INT8 Inference Using Quantization Aware Training with TensorRT*. <https://developer.nvidia.com/blog/achieving-fp32-accuracy-for-int8-inference-using-quantization-aware-training-with-tensorrt/>. July 2021.